研究论文
DOI：10.13998/j.cnki.issn1002-1248.23-0355

古农文语义检索模型构建及其应用研究

刘楠竹[1,2]，崔运鹏[1,2*]，王　末[1,2]
（1. 中国农业科学院农业信息研究所，北京　100081；2. 农业农村部 农业大数据重点实验室，北京　100081）

摘　要：[目的 / 意义]构建能实现以白话文作为查询，系统自动返回与输入最相关的古农文段落的语义检索模型，为学者提供更加便利的古代农业知识检索方式和古代农业知识溯源方式。[方法 / 过程]使用基于四库全书作为训练语料的 SikuBERT 作为基础模型，基于对比学习的方法，使用自建的古农文数据集对模型进行继续训练，得到能够支持使用白话文作为查询，返回与查询语义最相似的古农文段落的语义检索模型。[结果 / 结论]古农文语义检索模型的 Spearman 系数在测试集上的表现能够达到 86.51%，较基线模型在测试集上的表现 83.69%有一定程度的提升，在自建的古农文检索测试集上的召回情况(recall@k) 较基线模型有一定程度提升，模型在古农文上能够有比较好的检索效果。但受限于古农文训练语料规模，模型的训练效果还有很大提升空间。

关键词：古农文；语义检索；对比学习；模型构建；深度学习
中图分类号：TP391　　**文献标识码：**A　　**文章编号：**1002-1248（2023）07-0052-11

引用本文：刘楠竹,崔运鹏,王末. 古农文语义检索模型构建及其应用研究[J]. 农业图书情报学报,2023,35(7)：52-62.

1 引　言

中国是世界上著名的文明古国之一，拥有数千年未曾中断的农耕传统和农业历史，中国古农书作为中国传统农业经验知识、传统农业生产力和农业历史精髓的主要载体，未受到近代西方农学影响前中国人撰写的有关农业生产知识的著作，是中国珍贵的历史文化遗产。中国古农书中保存有大量资料有待发掘利用，传统作物栽培技术多半仍在现代农业中改造利用[1]，中国古农书中农业知识的价值并没有随着时代的前进而消失，对现代农业中出现的问题依然有切实指导作用。然而，中国古农书由古文写成，晦涩且无标点，艰深难读，普通大众只能望而却步。但古农书的价值需要传承和发扬，借助机器学习、深度学习等技术，对古农书中的知识进行挖掘，不仅能够方便学界进行研究，还可以向公众开放服务做知识普及，推动优秀文化遗产与当代社会发展的融合，实现古农书的跨越式发展。

2017 至 2022 年间，《关于实施中华优秀传统文化传承发展工程的意见》《中华人民共和国国民经济和

收稿日期：2023-05-29
基金项目：国家社会科学基金重大项目“中国古农书的搜集、整理与研究”（21&ZD332）
作者简介：刘楠竹（1991-），女，硕士研究生，研究方向为图书情报。王末（1987-），男，博士，副研究员，研究方向为农业信息技术、农业知识管理、数据挖掘技术研究
* 通信作者：崔运鹏（1972-），男，博士，研究员，研究方向为农业信息技术、农业知识管理、数据挖掘技术研究。Email：cuiyunpeng@caas.cn

社会发展第十四个五年规划和2035年远景目标纲要》《关于推进新时代古籍工作的意见》等多个文件都提出要“深入实施中华优秀传统文化传承发展工程”“加强文物和古籍保护研究利用”“提升中华文化影响力”“挖掘古籍时代价值、推进古籍数字化建设”。习近平总书记更是指出，要“让收藏在博物馆里的文物、陈列在广阔大地上的遗产、书写在古籍里的文字都活起来”。

为了解学者对中国古代农业书籍的使用需求，本研究使用文献调研的方法，在知网已发表论文中检索提到“古农书”“农业古籍”的文章，通过对检索结果进行分析找出古农书的使用者和使用需求。根据调查，对古农书内知识进行研究和直接引用的多为农史研究者、农业研究者及相关专业的学生。农史研究者在使用古农书时，通常会通过目录、版本、校勘、注释、考证、辨伪、辑佚等理论方法来分析、整理和研究古农书[2,3]，或是通过考究和利用古代有关农业遗存、配合古农书上的相关记录研究农业历史[4-7]，需要方便快速的根据主题、关键字或相关描述查询到尽量全的[8]相关古代农文书籍原文。农业研究者使用古农书一般为现代农业发展中出现了问题，需要对照现代农业的问题，从古农书中查询相关问题在历史上出现的缘由、时间、地点、解决方法、相关政策等信息，进行借鉴学习[9-11]，最终推出新的农业模式，也需要通过现代文的描述查找古农书中的相关内容。

目前中国专门的古农书数据库有以下几个：南京农业大学“中华农业文明网”，提供对《齐民要术》《农政全书》等16本农书的网络检索服务，可以通过目录关键字、书名、作者、朝代对古农书进行检索[12]；南京农业大学“中国农业遗产信息平台”提供200余种农业古籍的分类浏览检索和全文检索[13]；中国农业大学图书馆对几部经典农书如《齐民要术》《陈敷农书》等进行数字化处理，构建古农书图像数据库，并做主题标引，使读者可根据书名、作者、版本、年代对其进行检索；中国农业科学院图书馆“农业古籍珍藏及全文数字化研究与建设”项目，提供古籍书目数据库检索和重要农书的全文检索。可以看出，目前中国古农书的检索方式仍多为关键字检索或全文检索，想要获得更全面的检索结果，就需要通过多轮次的检索并且手动构建极为复杂的检索式，更适合对古籍、古农书有一定了解的学者使用，使用门槛较高，不能充分理解和适应读者需要。

为丰富中国古代农业书籍中文本内容的知识挖掘利用方式、提高知识的深度挖掘能力、激活古农书在新时代的传播力和生命力，本研究面向古农书的内容信息，即古农文，做语义检索模型构建及其应用研究。

语义检索指检索系统不再拘泥于用户问题的字面本身，而是能精准捕捉到用户问题背后的真正意图，并以此进行搜索，从而更准确地向用户返回最符合的结果。目前通过语义检索提高检索效果的实现方式主要有基于统计特征的语义检索、本体语义检索和向量语义检索，本文使用的语义检索方式为向量语义检索。

使用语义检索进行查询时，系统会先通过语义索引模型把用户查询和文档集合分别表示成可以反映其核心特征的向量，在高维向量空间中对它们进行索引。在这个向量空间内，语义相似的句子向量距离就比较近，从而可以使用余弦相似度、曼哈顿距离、欧氏距离等方法计算向量间的距离、度量用户查询和文档的相似程度，找出语义相似的句子，最终返回所有潜在的与问句相关的文档列表。

语义检索可以让使用者在未经过专门的检索工具培训情况下使用自然语言输入想搜索的内容，而不需构建检索式或使用特定语言来查询。若将语义检索应用到农业古籍领域，现代农业研究者、农史研究者和相关专业的学生等，都能够更方便的查询自己感兴趣的内容。

本研究基于BERT模型框架，利用《齐民要术》《氾胜之书》《管子地员篇》《吕氏春秋上农等四篇》《农桑辑要》《农政全书》《天工开物》《亢仓子农道第八》《补农书》《王祯农书》共10本农业古籍中的文本及其译文的平行语料，构建可实现从白话文到古农文的语义检索模型，即能够实现输入白话文、系统自动返回所有与输入句子相关的古农文段落的语义检索模型，为学者提供更加便利的古代农业知识检索方式和古代农业知识溯源方式。其使用场景为：学者在

DOI：10.13998/j.cnki.issn1002-1248.23-0355

写作时想使用一段古文做引证，但忘记了它的原文是什么，只记得大概的意思，或是在看一本古籍时，其中对于某一个方法、理念，好像以前在哪本古籍中看到过，想查一下该方法、概念的内容沿袭情况时，可以使用语义检索系统，直接输入现代文的描述，系统自动返回与描述最相关的古文。

2 文献综述

2.1 语义检索

在工业上由于需要从大型数据库中返回一组相关文档，为了平衡搜索效率和效果，语义检索通常采用多阶段的排名策略，即“召回和重排”策略[14,15]。召回和重排阶段都是通过对查询和文档的相关性进行评估返回文档，但根据其目的的不同通常使用不同的模型[16]。召回阶段的目标是从庞大的文档库中召回所有潜在的相关文档，然后传递给重排阶段。在重排阶段，由于需要考虑的文档数量较少，通常会采用更复杂的排序模型构成重新排序器，使用 1 个或多个重新排序器对已召回的文档进行排序，每个重新排序器接收前一个重新排序器生成的排序列表，依次其进行重排，最后将最终的重排结果返回给用户。因此召回阶段通常优先考虑效率和高召回率，重排阶段更多的考虑有效性。

本文主要对古农文语义检索第一阶段召回模型进行研究，为方便称呼，后续直接称其为古农文语义检索模型。

语义检索的发展分为 3 个阶段：基于术语的语义检索、基于特征表示阶段和神经语义检索。

语义检索最开始是将查询和文档分别用离散的词袋（BOW）表示，再利用倒置索引技术来管理大规模文档，如 BM25（术语匹配 +TF-IDF 权重），这种基于术语的检索模式由于其简单的逻辑和强大的索引，得到了非常好的召回效果[17,18]。但由于独立性假设，它们可能出现词汇不匹配[19,20]，并且由于没有考虑术语出现的序列，它们可能不能很好的捕捉到文档的语义[21]。

为了得到更好的检索效果，学者们进行了大量的工作，如使用查询拓展[22-25]、文档拓展[26-28]、术语依赖性模型[29]、主题模型、基于信息检索的翻译模型等。但以上方法仍然处在词袋的表征范围内，依赖手工制作的特征来建立表示函数，旨在用从外部资源或集合本身提取的语义单元来改进经典的词袋表示，只能捕捉到浅层的句法和语义信息，未能突破其局限性[30]。

2013 年之后，随着特征表示学习方法的发展，词嵌入技术[31-33]逐渐被应用到语义检索中。与离散的符号表示不同，词嵌入是一种密集的表示，可以一定程度上缓解词汇的错误匹配问题。2016 年之后，随着深度学习技术的发展，学者们开始在传统的离散符号表示范式中改进文档表示[34,35]，或在稀疏表示和密集表示范式中直接形成一系列新的语义检索模型[36-39]，一般称之为神经语义检索方法。神经检索方式能够通过神经网络建立表示函数和评分函数的检索方式[40,41]，使用词嵌入技术来捕获单词的语义属性，并以端到端的方式从数据中学习深层语义和复杂的互动关系。

2.2 相关预训练模型

随着深度学习的发展，模型参数的数量迅速增加，需要更大的数据集来完全训练模型参数并防止过拟合[42]，但构建大规模的标注数据集是一个巨大的挑战，预训练模型可以从容易获得的大规模无标注数据上学习到通用、良好的语言表示和更好的初始化参数[42]，然后将这些表示形式用于其他任务。由于古文标注资源的稀缺性，在做相关方面的研究时，需要使用预训练模型来增强模型在低资源环境下的文本处理效果。

目前针对英文和现代文通用领域的语义索引模型已经有了很多的研究，在工程上也取得了比较好的效果。但是在古文领域，由于缺乏大规模纯净的古文及其译文数据，构建古文标注训练集成本高昂，对数据标注人员具有较高要求[42]，古文领域的预训练语言模型很少。

目前面向古文的预训练语言模型只有北京理工大学阎覃等的 GuwenBERT、南京农业大学王东波等的 SikuBERT 和 SikuRoBERTa、WANG 的 Bert-Antient-Chinese[44]，这 4 个模型都是以 BERT 类模型作为基线模型，使用不同的训练数据进行训练得到的预训练语

言模型。在模型训练上，GuwenBERT 是基于继续训练技术，在中文 RoBerta 的基础上，使用殆知阁古文数据（包含 15 694 本古文书籍，字符数 1.7B，所有繁体字均经过简体转换处理）进行迁移学习训练出的预训练语言模型，能够在简体中文下获得较好的古文处理性能；SikuBERT 和 SikuRoBERTa 是在中文 BERT 和中文 RoBERTa 基础上，基于领域适应训练的思想，使用繁体《四库全书》全文语料（字数达 536 097 588 个，数据集内的汉字均为繁体中文）训练出的面向古文自动处理领域的预训练语言模型，该模型更适用于繁体古籍处理；Bert-Antient-Chinese 是在中文 BERT 的基础上，基于领域适应训练的思想，结合古文语料进行继续训练得到的面向古文自动处理领域的预训练模型，训练时使用涵盖了从部、道部、佛部、集部、儒部、诗部、史部、医部、艺部、易部、子部作为训练集，训练集规模约为《四库全书》6 倍大的语料进行继续训练得到的拥有更大词表的预训练语言模型（词表大小为 38 208，SikuBERT/SikuRoBERTa 为 29 791），能够同时适用于繁体和简体。

支持通过向量语义相似度进行从白话文到古文的语义检索的模型只有南京农业大学的 BTfhBER 和 ZHANG 的 XLsearch-cross-lang-search-zh-vs-classicical-cn。在模型训练上，BTfhBER 是在中文 BERT 的基础上，基于二十四史古白平行语料继续训练的古白跨语言预训练模型；XLsearch-cross-lang-search-zh-vs-classicical-cn 是在 BERT-base-Chinese 的基础上，使用约 90 多万句古白平行句对进行训练，得到的古白跨语言模型。二者在从白话文到古文的语义文本相似度任务上可以获得较好的效果，但在进行从白话文到古农书内文本的语义检索任务时，效果仍有欠缺。

3　数据与方法

3.1　数据源简介

本实验使用自行构建的古农文语义检索数据集，其中共有古白平行语料数据 9 542 对，其中正例 4 771 对，通过随机采样生成负例 4 771 对，共含汉字 1 535 514 个，所有汉字均为繁体字。将句对按 8:1:1 的比例分成训练集、验证集、测试集。

模型输入时，使用（x_i，x_i^+，x_i^-）的形式，向模型输入原句、正例和负例，使得模型在保证正例间相似度的同时将负例的距离推远。最终形成训练集中含三元组数据 3 808 条，测试集和验证集中各含句对 952 对。

以上数据共来源于 10 本农业古籍原文及其译文：《齐民要术》《氾胜之书》《管子地员篇》《吕氏春秋上农等四篇》《农桑辑要》《农政全书》《天工开物》《亢仓子农道第八》《补农书》《王祯农书》。其中，古农文文本数据来源于殆之阁和国学梦两个国学经典网站，对应的译文分别使用梁乐和许蕻翻译、巴蜀书社 1995 年出版的《齐民要术白话全译》，韦占彬、张春花的《农政全书译文》，陈恒力、王达的《补农书校释》，缪启愉、缪桂龙的《东鲁王氏农书译注》，国学荟《天工开物》译文，查字典诗词网《农桑辑要》译文，豆瓣如是《氾胜之书试译》，华韵国学网《吕氏春秋》上农等 4 篇译文，书摘天下《管子》地员篇译文，天蚨园《亢仓子农道第八》。

最后，为了检测模型在召回任务上的性能，本实验还使用基于汉语古典文本数据库 scripta-sinica 进行微调的古汉语问答模型 Bloom 春华，对除上述 10 本农书之外的其他农书中随机抽取的段落进行翻译，构建出用于进行模型召回效果评价的数据集，其中共含古白句对 477 对。

3.2　古农文预训练模型构建

本研究构建了一个古农文语义检索模型，能够实现使用白话文作为查询，检索出数据库中与查询语义相似度最高的 k 个古农文段落的功能。实验共分为 4 个部分：语料预处理、模型训练、模型效果评价和语义检索任务测试。

实验先根据 10 本古农书文本数据及其译文数据的情况进行清洗和数据对齐，构建出古白平行语料，按照“8:1:1”划分训练集、验证集和测试集。

DOI：10.13998/j.cnki.issn1002-1248.23-0355

模型训练阶段，根据预实验结果对训练参数进行调整，使用 Pytorch 版的 SikuBERT 模型和 SimCSE 有监督训练框架，在训练集和验证集上完成模型的训练。在效果评价阶段，在测试集上使用文本相似度任务判断模型训练效果，再通过语义检索任务分析模型检索性能。

3.2.1 训练模型选取

BERT（Bidirectional Encoder Representations from Transformers）[45]是 2018 年谷歌提出的面向自然语言处理任务的自监督预训练语言模型，它使用 Transformer 的双向编码器结构作为特征提取器，为了实现文本的双向建模，BERT 采用一种类似完形填空的做法来实现基于自编码的预训练任务，即为掩码语言模型（MLM，Masked Language Model）。MLM 在预训练任务时将输入文本中的部分单词 Mask 并还原为原单词用以避免双向语言模型带来的信息泄露问题，使得模型通过被掩码词周围的上下文信息来还原掩码位置的词，从而学习上下文敏感的文本表示。为了学习两断文本之间的关联，BERT 还通过下一句预测任务（NSP，Next Sentence Prediction），即通过判断句子 B 是否是句子 A 的下一个句子来构建两段文本之间的关系[46]。BERT 通过掩码语言模型的方法训练词的语义理解能力，下一句预测的方法训练句子之间的理解能力，使得它能够很好的支持那些涉及句子间语义联系判断、需要对文本进行深层语义理解、需要分析句子语义信息的下游任务[47]。而 BERT 的“预训练 - 微调”训练方法也使得只需要对模型的高层参数进行调整，就能让模型适应不同的下游任务。

谷歌 2018 年发布的基础 BERT 主要适用于英文，为了拓展其适用范围，谷歌后续又发布了使用中文维基百科训练的面向中文的预训练语言模型 BERT-Base-Chinese。

SikuBERT[43]是在中文 BERT 的基础上，基于领域适应训练的思想，使用繁体《四库全书》全文语料（字数达 536 097 588 个，数据集内的汉字均为繁体中文）训练出的面向古文自动处理领域的预训练语言模型。

3.2.2 模型训练方法

原生 BERT 在进行语义文本相似性等句子对回归任务时并没有计算独立的句子嵌入，而是通过交叉编码器将两个句子拼接成一个序列后传递给 Transformer 来预测目标，这种方式使得它在时间和计算量上都会产生巨大的开销，大量的实验也证明了 BERT 开箱即用地将句子映射到一个向量空间并不适合用于常见的相似度度量（如余弦相似度），直接使用原生 BERT 会产生很糟的句子嵌入[48]，在语义相似度任务上表现并不好。

SimCSE[49]是一个简单的对比学习框架，通过拉近语义上接近的句子、推远语义不相似的句子来增强句子嵌入的学习效率，能够极大的提高在语义文本相似度上句向量的质量。训练时，除了将给出的矛盾数据作为困难负例，SimCSE 还将批次中其它句子作为负例，使得在减小相似样本间距离的同时，增加不相似样本间的距离，提高向量表征。例如：假如一个批次中有 N 个三元组句对，每个句子就有 1 个正例和 N 个负例。在数据扩充的同时，SimCSE 还通过在训练过程中还通过正例对齐性（Alignment）和空间一致性（Uniformity）来衡量词嵌入的学习质量。正例对其性能够计算句对嵌入的预期距离，空间一致性能够衡量嵌入均匀分布的程度。有文章通过实证分析，对齐性和一致性与对比学习的目标一致，提高空间一致性能够缓解 BERT 的各向异性。一般来说，具有更好的正例对齐性和空间一致性的模型可以获得更好的性能。

本文分别使用 BERT-Base-Chinese 和 SikuBERT 作为基础模型，使用 SimCSE 有监督训练框架，在自建的古农文数据集上进行古农文语义检索模型的训练。

3.2.3 模型效果评价指标

斯皮尔曼等级相关系数（Spearman's Rank Correlation Coefficient，Spearman 相关系数 ρ）是衡量两个变量的依赖性的非参数指标。它利用单调方程评价两个统计变量的相关性。如果数据中没有重复值，并且当两个变量完全单调相关时，斯皮尔曼相关系数则为 +1 或−1。即斯皮尔曼相关系数是衡量排名而不是实际分数的，更适合评估句子嵌入。

由于本项目数据集是把句子对进行 0、1 打分来区分是否相似的，所以在模型效果评价阶段，评测指标采用斯皮尔曼等级相关系数，即给定句对，模型通过计算句子嵌入的余弦相似性和黄金标签之间的斯皮尔曼秩相关性来判断两个句子的语义是否相同。对于样本容量为 n 的样本，n 个原始数据 X、Y 被转换成等级数据 x、y，相关系数为斯皮尔曼相关系数 ρ，得分越高说明相关性越高。

指标计算公式如下：

$$\rho=\frac{\sum_i(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2\sum(y_i-\bar{y})^2}}$$

Recall@k 召回率，指前 TopK 结果中检索出的相关结果数和库中所有的相关结果数的比率，衡量的是检索系统的查全率。

$$Recall@k=\frac{true_positives@k}{all_positive}$$

其中 true positices@k 表示 k 个预测结果中正例的数量，all_positive 表示全库中所有的正例数。

3.2.4　模型参数设置

古农文语义检索模型训练时，使用的参数如表 1 所示。

表 1　主要超参数设置

Table 1　Main hyperparameter settings

超参数	解释	值
max_seq_length	最大输入序列长度	512
train_batch_size	每个批次训练数据大小	16
learning_rate	学习率	2e-5
warmup_steps	预热学习步数	训练集的 10%
num_epochs	训练周期	3

4　实验结果及分析

4.1　Spearman 系数

直接使用 SikuBERT 和 BERT-Base-Chinese 对验证集进行模型语义相似度预测时，其 Spearman 系数分别为 83.69%和 69.52%。

训练后的 SimCSE-BERT-Base-Chinese 和 SimCSE-SikuBERT 在古农文模型在验证集上的 Spearman 系数为 86.14%和 86.51%，相比原始模型在测试集上的表现有较大提升，说明模型学习到了相关古农文知识。

4.2　模型效果

从表 2 可以看出，训练后的 SimCSE-BERT-Base-Chinese 和 SimCSE-SikuBERT 在使用大语言模型 Bloom 春华进行拓展的测试集上，召回效果要强于直接使用基础模型。

由于 SimCSE-SikuBERT 的整体效果最好，所以本实验选择 SimCSE-SikuBERT 作为古农文语义检索模型。

4.3　检索效果

本实验使用 SikuBERT 和训练后的古农文语义检索模型进行小规模的检索实验。

本实验采用双编码器结构，将查询和数据库中的段落分别使用相同的编码器（语义检索模型）进行独立编码，得到稠密表示向量。由于语料库中数据较少，使用余弦相似度计算问题和语料库中所有段落的相似度，返回得分最高的 2 个段落。

在语义检索文档数据方面，本项目根据王毓瑚

表 2　模型召回效果

Table 2　Model recall effect

模型	Recall@1/%	Recall@5/%	Recall@10/%
SikuBERT mean	81.55	90.15	92.03
BERT-Base-Chinese mean	54.93	70.02	76.73
SimCSE-BERT-Base-Chinese	96.02	98.56	98.96
SimCSE-SikuBERT	96.44	98.74	99.00

DOI：10.13998/j.cnki.issn1002-1248.23-0355

《中国农学书录》记载的中国古农书轶存情况，对中国古农书的数字化文本进行搜寻，共搜集到 108 本农书。为了方便使用搜集到的古农文文本内容进行模型的训练和后续的语义检索，本项目对搜集到古农书文本，即古农文，进行整理、清洗、分段、去重，最终得到 14 133 条古农文数据，将其作为古农文语义检索模型进行语义检索任务的文档数据。通过知网参考文献中有《齐民要术》《王祯农书》《补农书》等农业古籍的论文，查看作者引用目的，准备一些用现代文描述的古代农业技术知识作为查询，观察模型在古农文语义检索中的性能。

实验中使用的皆为中文繁体字，为方便查看，使用 OpenCC 将查询和检索结果转换成中文简体。

从表 3 可以看出，使用训练过的语义检索模型，检索出的句子质量更高，并且除了现代文对应的古文，还能够检索出其它意思相近的古文，起到知识溯源的效果。

5 结　语

实验结果表明，基于 SimCSE 框架对基础古文语言模型 SikuBERT 进行训练，能够使模型学习到单词的上下文信息，生成句子级的嵌入表达，提升语义检索速度，且其检索结果较基线模型有一定程度的提升，能够有效提升在古农文上的语义检索效果，验证了双语和反翻译语料库可以为语义相似性学习提供有用的监督，基于 BERT 的“预训练 - 微调”训练方法在 SimCSE 结构上是可行的。

受限于古农文语料数据数量及质量，使得目前古农文语义检索模型的效果依然不如预期，训练时使用的古农文古白句对在词汇重叠上往往比较多，这会影响模型的学习效率，使得模型在用于搜索时更容易寻找句面相似而非语义相似的句子。同时，本项目训练时使用的古白数据集是没有困难负例的，构造时直接

表 3　语义检索试验结果

Table 3 Semantic retrieval experiment results

	SikuBERT mean	SimCSE-SikuBERT
Query1：在豆叶落尽的时候要全部收割，或者在豆角青黄相间的时候将植株拔出，扎拢倒置，这样成熟的小豆不受天气影响，颗粒饱满		
#1	下接力，须在处暑后，苗倣胎时，在苗色正黄之时。如苗色不黄，断不可下接力；到底不黄，到底东可下也。若苗茂密，度其力短，俟抽穗之后，每亩下饼三斗，自足接其力。切不可未黄先下，致好苗而无好稻	夫收割之法，待其可收则刈。豆角三青两黄，拔而倒竖笼从之，则生熟皆均，不畏严霜，从本至末，全无秕减
#2	夫收割之法，待其可收则刈。豆角三青两黄，拔而倒竖笼从之，则生熟皆均，不畏严霜，从本至末，全无秕减	叶落尽，则刈之。叶未尽者，难治而易湿也。豆角三青两黄，拔而倒竖笼从之，生者均熟，不畏严霜，从本至末，全无秕减，乃胜刈者
Query2：为预防气候变化，应既种早谷，亦种晚谷，不宜只种一种；闰年节季稍晚，应当迟种；在正常年分，应以早种为佳，早种量应超过晚种量的一倍		
#1	芒种有二义：郑元谓有芒之种。若今黄穋谷是也。一谓待芒种节过乃种。今人占候，夏至小满至芒种节，则大水已过，然后以黄穋谷种之于湖田。然则有芒之种与芒种节候二义，可并用也。黄穋谷自初种以至收刈，不过六七十日，亦可以避水溢之患	凡田欲早晚相杂，防岁道有所宜。有闰之岁，节气近后，宜晚田。然大率欲早，早田倍多于晚田
#2	春大豆，次稙谷之后。二月中旬为上时，一亩用子八升；三月上旬为中时，亩用子一斗；四月上旬为下时，亩用子一斗二升。岁宜晚者，五六月亦得；然时晚则种子当稍加，地不求熟故也。尤当及时锄治，使之叶蔽其根，庶不畏旱	防歲道有所宜。有閏之歲，節氣近後宜晚田，然大率欲早，早田倍多於晚

使用其它古农文的译文作为负例，这也会导致训练结果无法进一步提升。语义检索在构建段落向量时，向量受段落组织方式影响较大，合适的分段方式可以提高段落被正确检索出的概率，若想获得更好的检索结果还需要更合适的古农文信息组织方式。

下一步工作将继续探索合适的训练范式，充分利用模型，同时探寻将已有的古农史知识融合到模型训练中去，达到更好的古农文语义检索效果。

参考文献：

[1] 张波. 农史研究法[M]. 咸阳: 西北农林科技大学出版社, 2019.
ZHANG B. Agricultural history research method [M]. Xianyang: Northwest A&F University Press, 2019.

[2] 葛小寒. 文献、史料与知识——古农书研究的范式及其转向[J]. 中国农史, 2019, 38(2): 12–25.
GE X H. Text, history date and knowledge – The paradigms of ancient agricultural books' research in agricultural history of China[J]. Agricultural history of China, 2019, 38(2): 12–25.

[3] 何凡能, 李柯, 刘浩龙. 历史时期气候变化对中国古代农业影响研究的若干进展[J]. 地理研究, 2010, 29(12): 2289–2297.
HE F N, LI K, LIU H L. The influence of historical climate change on agriculture in ancient China[J]. Geographical research, 2010, 29(12): 2289–2297.

[4] 曾雄生. 也释“白田”兼“水田”——与辛德勇先生商榷[J]. 自然科学史研究, 2012, 31(2): 201–208.
ZENG X S. An alternative interpretation of Baitian (white field) and Shuitian (water field): Discussion with Mr. Xin Deyong[J]. Studies in the history of natural sciences, 2012, 31(2): 201–208.

[5] TANG M, WANG X, HOU K, et al. Carbon and nitrogen stable isotope of the human bones from the Xiaonanzhuang cemetery, Jinzhong, Shanxi: A preliminary study on the expansion of wheat in ancient Shanxi, China[J]. Acta anthropologica sinica, 2018, 37(2): 318–30.

[6] 刘志国, 徐旺生.《齐民要术》的盐史信息考探[J]. 中国科技史杂志, 2021, 42(1): 91–99.
LIU Z G, XU W S. The information on salt history in the qimin Yaoshu[J]. The Chinese journal for the history of science and technology, 2021, 42(1): 91–99.

[7] ZHOU X Y, ZHU L, SPENGLER R N, et al. Water management and wheat yields in ancient China: Carbon isotope discrimination of archaeological wheat grains[J]. The holocene, 2021, 31(2): 285–293.

[8] CHEN S C. Exploring the use of electronic resources by humanities scholars during the research process[J]. Electron libr, 2019, 37: 240–254.

[9] WANG S Y, CUI D A, LV Y N, et al. Cangpu oral liquid as a possible alternative to antibiotics for the control of undifferentiated calf diarrhea[J]. Frontiers in veterinary science, 2022, 9: 879857.

[10] XIA X Y, LIN Z C, SHAO K P, et al. Combination of white tea and peppermint demonstrated synergistic antibacterial and anti–inflammatory activities[J]. Journal of the science of food and agriculture, 2021, 101(6): 2500–2510.

[11] WANG N, LIU X, LI J G, et al. Antibacterial mechanism of the synergistic combination between streptomycin and alcohol extracts from the Chimonanthus salicifolius S. Y. Hu. leaves[J]. Journal of ethnopharmacology, 2020, 250: 112467.

[12] 李明杰, 陈梦石, 孟彬. 中国古代科技文献整理出版七十年回望(1949–2019)[J]. 出版科学, 2019, 27(5): 22–29.
LI M J, CHEN M S, MENG B. Review on the collation of ancient Chinese scientific and technological documents in the past 70 years[J]. Publishing journal, 2019, 27(5): 22–29.

[13] 曹玲, 常娥, 薛春香. 农史研究的新工具——中国农业遗产信息平台的设计与构建[J]. 中国农史, 2006, 25(1): 127–133.
CAO L, CHANG E, XUE C X. A new tool of agricultural history research – Design and construction of "agricultural inheritance information database"[J]. Agricultural history of China, 2006, 25(1): 127–133.

[14] LIU S C, XIAO F, OU W W, et al. Cascade ranking for operational e–commerce search[J]. arXiv: 1706.02093, 2017.

[15] PEDERSEN J. Query understanding at being[R]. Invited Talk: SIGIR, 2010.

[16] FAN Y X, XIE X H, CAI Y Q, et al. Pre–training methods in information retrieval[M]. Beijing: Now Publishers, 2022.

[17] CHEN R C, GALLAGHER L, BLANCO R, et al. Efficient cost–aware cascade ranking in multi–stage retrieval[C]// Proceedings of the 40th International ACM SIGIR Conference on Research and Development

DOI：10.13998/j.cnki.issn1002-1248.23-0355

in Information Retrieval. New York: ACM, 2017: 445-454.

[18] LIANG D, XU P, SHAKERI S, et al. Embedding-based zero-shot retrieval through query generation[J]. arXiv preprint arXiv:200910 270, 2020.

[19] FURNAS G W, LANDAUER T K, GOMEZ L M, et al. The vocabulary problem in human-system communication[J]. Communications of the ACM, 1987, 30(11): 964-971.

[20] ZHAO L, CALLAN J. Term necessity prediction[C]// Proceedings of the 19th ACM international conference on Information and knowledge management. New York: ACM, 2010: 259-268.

[21] LI H, XU J. Semantic matching in search[J]. Foundations and trends® in information retrieval, 2014, 7(5): 343-469.

[22] LAVRENKO V, CROFT W B. Relevance based language models[C]// Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2001: 120-127.

[23] LESK M E. Word-word associations in document retrieval systems[J]. American documentation, 1969, 20(1): 27-38.

[24] QIU Y G, FREI H P. Concept based query expansion[C]// Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 1993: 160-169.

[25] XU J X, CROFT W B. Quary expansion using local and global document analysis[J]. ACM SIGIR forum, 2017, 51(2): 168-175.

[26] AGIRRE E, ARREGI X, OTEGI A. Document expansion based on WordNet for robust IR[C]. Posters: In Proceedings of COLING 2010, 2010: 9-17.

[27] EFRON M, ORGANISCIAK P, FENLON K. Improving retrieval of short texts through document expansion[C]// Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2012: 911-920.

[28] LIU X Y, CROFT W B. Cluster-based retrieval using language models[C]// Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2004: 186-193.

[29] GAO J F, NIE J Y, WU G Y, et al. Dependence language model for information retrieval [C]// Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2004: 170-177.

[30] GUO J F, CAI Y Q, FAN Y X, et al. Semantic models for the first-stage retrieval: A comprehensive review[J]. ACM transactions on information systems, 40(4)1-42.

[31] BOJANOWSKI P, GRAVE E, JOULIN A, et al. Enriching word vectors with subword information[J]. Transactions of the association for computational linguistics, 2017, 5: 135-146.

[32] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[J]. arXiv: 1310.4546, 2013.

[33] PENNINGTON J, SOCHER R, MANNING C. Glove: Global vectors for word representation [C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg, PA, USA: Association for Computational Linguistics, 2014.

[34] DAI Z Y, CALLAN J. Context-aware sentence/passage term importance estimation for first stage retrieval[J]. arXiv: 1910.10687, 2019.

[35] NOGUEIRA R, YANG W, LIN J, et al. Document expansion by query prediction[J]. ArXiv: 1904.08375, 2019.

[36] GILLICK D, PRESTA A, TOMAR G S. End-to-end retrieval in continuous space[J]. arXiv: 1811.08008, 2018.

[37] JANG K R, KANG J M, HONG G, et al. UHD-BERT: Bucketed ultra-high dimensional sparse representations for full ranking[J]. 2arXiv: 2104.07198, 2021.

[38] KHATTAB O, ZAHARIA M. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT[C]// Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 2020: 39-48.

[39] ZAMANI H, DEHGHANI M, CROFT W B, et al. From neural re-ranking to neural ranking: Learning a sparse representation for inverted indexing[C]// Proceedings of the 27th ACM International Conference on Information and Knowledge Management. New York: ACM, 2018: 497-506.

[40] BARONI M, DINU G, KRUSZEWSKI G. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors[C]// Proceedings of the 52nd Annual Meeting of the Associ-

ation for Computational Linguistics (Volume 1: Long Papers). Stroudsburg, PA, USA: Association for Computational Linguistics, 2014.

[41] BENGIO Y, DUCHARME R, VINCENT P, et al. A neural probabilistic language model[J]. J Mach learn res, 2003, 3: 1137–1155.

[42] QIU X P, SUN T X, XU Y G, et al. Pre–trained models for natural language processing: A survey[J]. Science China technological sciences, 2020, 63(10): 1872–1897.

[43] 王东波, 刘畅, 朱子赫, 等. SikuBERT 与 SikuRoBERTa：面向数字人文的《四库全书》预训练模型构建及应用研究[J]. 图书馆论坛, 2022, 42(6): 31–43.
WANG D B, LIU C, ZHU Z H, et al. Construction and application of pre–trained models of siku Quanshu in orientation to digital humanities[J]. Library tribune, 2022, 42(6): 31–43.

[44] WANG P Y, REN Z C. The uncertainty–based retrieval framework for ancient Chinese CWS and POS[C]. Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages, 2022: 164–8.

[45] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre–training of deep bidirectional transformers for language understanding[J]. arXiv: 1810.04805, 2018.

[46] 车万翔, 郭江, 崔一鸣. 自然语言处理: 基于预训练模型的方法[M]. 北京: 电子工业出版社, 2021.
CHE W X, GUO J, CUI Y M. Natural language processing[M]. Beijing: Publishing House of Electronics Industry, 2021.

[47] 邵浩, 刘一烽. 预训练语言模型[M]. 北京: 电子工业出版社, 2021.
SHAO H, LIU Y F. Pre–training language model[M]. Beijing: Publishing House of Electronics Industry, 2021.

[48] REIMERS N, GUREVYCH I. Sentence–BERT: Sentence embeddings using Siamese BERT–networks[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP–IJCNLP). Stroudsburg, PA, USA: Association for Computational Linguistics, 2019.

[49] GAO T Y, YAO X C, CHEN D Q. SimCSE: Simple contrastive learning of sentence embeddings[J]. arXiv: 2104.08821, 2021.

Construction and Application of Semantic Retrieval Model for Ancient Agricultural Literature

LIU Nanzhu[1,2], CUI Yunpeng[1,2*], WANG Mo[1,2]

（1. Institute of Agricultural Information, Chinese Academy of Agricultural Sciences, Beijing 100081;

2. Key Laboratory of Agricultural Big Data, Ministry of Agriculture and Rural Affairs, Beijing 100081）

Abstract: [Purpose/Significance] The ancient Chinese agricultural books are the main carrier of traditional agricultural experience, and represent the productivity and the essence of agricultural history in China. The value of agricultural knowledge in them has not disappeared with the progress of the times, and still has practical guidance for the problems that arise in modern agriculture. However, the ancient Chinese agricultural books are written in ancient Chinese, which are obscure and without punctuation, making them difficult to use. Semantic retrieval is a retrieval method that automatically queries and extracts relevant information from information sources at the semantic level. It can accurately capture the true intention behind user problems and conduct searches based on it, and thereby it is

DOI：10.13998/j.cnki.issn1002-1248.23-0355

capable of returning more accurate and the most consistent results to users. However, currently most relevant research only focuses on major languages, and there is insufficient research on sentence embedding in ancient Chinese prose. In order to fill the gap in the field and provide scholars with more convenient methods for retrieving ancient agricultural knowledge and tracing ancient agricultural knowledge, this study is based on comparative learning methods to construct a semantic retrieval model that can automatically return the most relevant ancient agricultural paragraph with input, using vernacular Chinese as the query. [Method/Process] SikuBERT, which is based on Siku Quanshu as the training corpus, is used as the basic model. Based on the method of comparative learning, the model is continued to be trained using the self-built ancient agricultural dataset, and a semantic retrieval model that can support the use of vernacular as a query and return the ancient agricultural paragraphs most similar to the query semantics is obtained. [Results/Conclusions] The Spearman coefficient of the ancient agricultural text semantic retrieval model can achieve 86.51% performance on the test set, which is a certain degree of improvement compared to the baseline model's 83.69% performance on the test set. The recall situation on the self built ancient agricultural literature retrieval test set has been improved to a certain extent compared to the baseline model, and the model can have good retrieval results on ancient agricultural literature. However, semantic retrieval models usually require relevant semantic similarity datasets or semantic matching datasets for training. Due to the lack of large-scale and pure ancient Chinese data in the field of ancient agricultural literature, and the high cost of constructing relevant datasets requiring personnel with high-standard relevant professional qualifications, this experiment used a self-built dataset for training, which is limited by the quantity and quality of ancient agricultural language corpus data. The current semantic retrieval model for ancient agricultural literature is still not as effective as expected. In the future, we will search for training methods suitable for small samples, such as transfer learning based on cross language pre-training models to improve the retrieval performance.

Keywords: ancient agricultural script; semantic retrieval; comparative learning; model building; deep learning